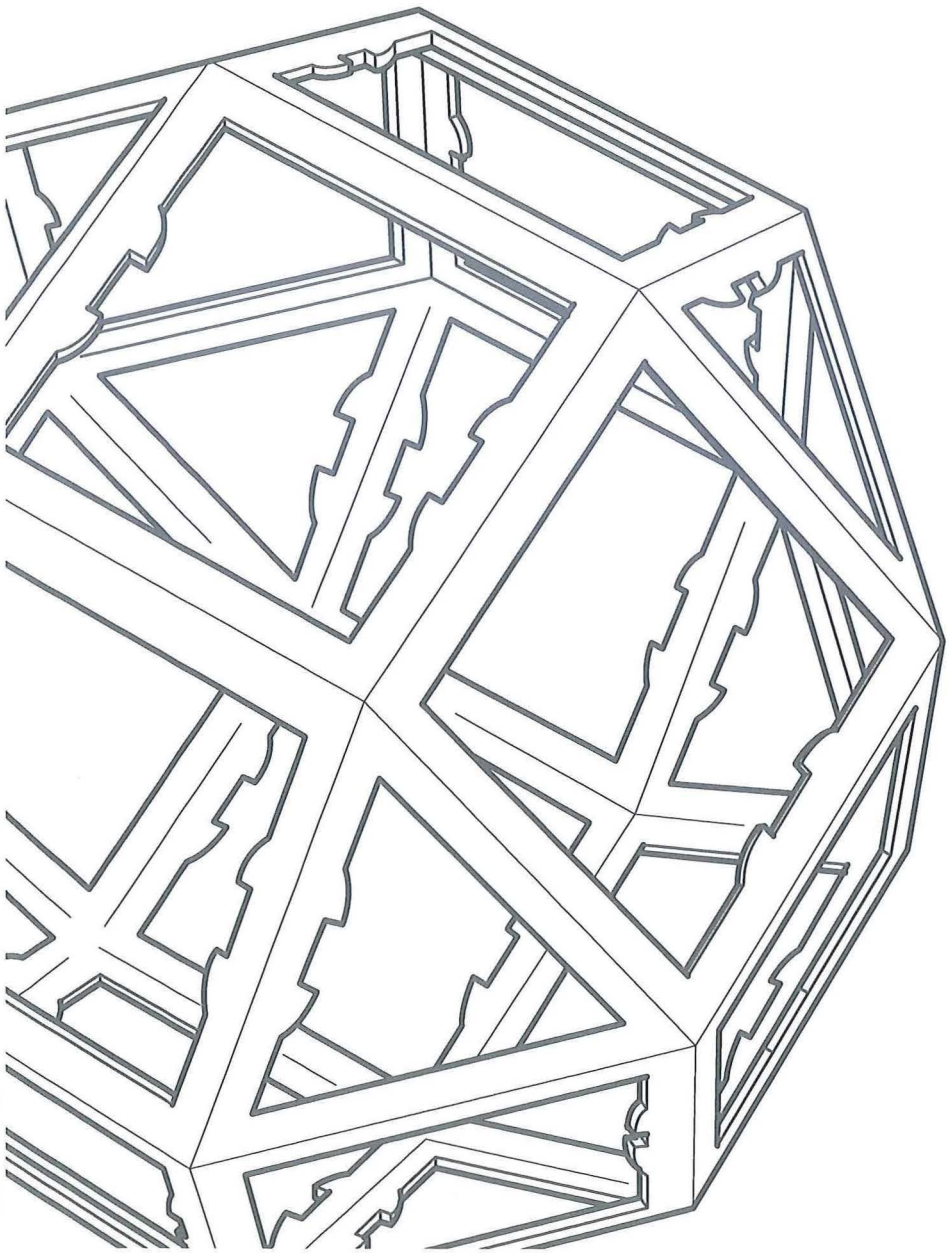
Department of Mathematics and Statistics
College of Engineering

UNIVERSITY OF CANTERBURY
Te Whare Wānanga o Waitaha
CHRISTCHURCH NEW ZEALAND

Summer Research Project

Visualizing Coalescent Theory

by Miriam Hodge



08

Summer Research Project

Visualizing Coalescent Theory

Miriam Hodge

Supervisor: Raazesh Sainudiin

8 February 2008

1 Abstract

Coalescent theory is a field of statistical genetics that can offer insight to molecular evolutionary biologists. In order to apply Coalescent Theory in their field studies, biologists must understand the coalescent models. This study is aimed at providing interactive visualizations to supplement algorithmic learning of the basic coalescent models.

2 Introduction

Coalescent Theory is built around three insights, "The first is the idea of tracing the ancestry of a gene backward in time and building up the family tree of the genes (at a particular locus) in a population sample back to the point at which they have a single common ancestor. ... [T]he second insight,

that for a large class of demographic models ... the stochastic structure of the genealogy does not depend on the detail of the reproductive mechanism. Finally, for such models the effect of mutation is statistically independent of the genealogy[5]."

This study will focus on the first two insights. We first review two demographic models with differing reproductive methods. We will then visually compare the structure of their genealogies. Finally, we build ancestries backward in time.

In order to make the resulting mathematics tractable the demographic models are very simple. The models presented in this study have no recombination, no natural selection and no structure to the population. Simulating these models under different conditions and generating visualizations allows us to reinforce the model's mathematical properties. The aim of the study is to develop same intuition about how these models change over time.

3 Definitions

We begin by defining some biological terms that may not be familiar to all readers.

Haploid: An organism with one set of chromosomes.

Diploid: An organism with two sets of chromosomes.

Generation Zero: The starting population for a simulation.

Asexual Reproduction: Genetic material of offspring is an exact copy of the parent.

Sexual Reproduction: Genetic material of offspring is a combination of multiple parents.

DNA: Deoxyribonucleic acid, the mechanism for transmitting genetic information. For our purposes it can be viewed as an ordered string comprised of the four DNA Bases.

DNA Bases: Adenine(A), Guanine(G), Cytosine(C), and Thymine(T)

Segregating Sites: Sites found when comparing DNA strings that are not identical for all individuals.

Site: The location of a single DNA base in an Allele

Alleles: An ordered string of bases from a given position on a string of DNA

MRCA: Most Recent Common Ancestor, the most recent individual that has all individuals of interest as direct descendants.

4 Data Visualization

Data Visualization converts numeric information into a graphic format. One of the things that visualization can do is help us understand the relative size of numbers. For example, the following visualization conveys the difference between 1 and 100. The blue box is one-hundred times larger than the green box.

"The greatest possibilities of visual display lie in vividness and inescapability of the intended message. A visual display can stop your mental flow in its tracks and make you think. A visual display can force you to notice

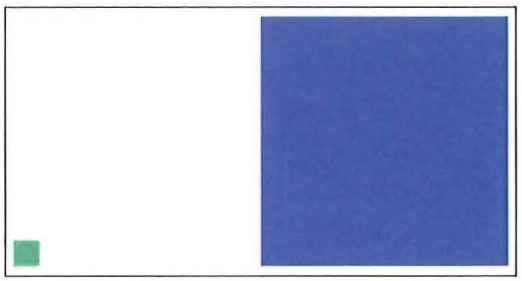


Figure 1: Relative Size Visualization

what you never expected to see. ("Why, that scatter diagram has a hole in the middle!")[6]".

Visualization aid our understanding of data, but they can not replace mathematical rigor. We sometimes detect patterns that do not really exist[3, pg. 45].

The visualizations of population models presented in this paper are built on two matrices of data. The first one is the image matrix it is a g x N matrix, where g is the number of generations and N is the total population size. Each element in the matrix is represented by a pixel in the resulting image. Each pixel in the image represents an individual at a given time point. As you move across the matrix from left to right you are looking at different individuals at the same time point. As you move down time is advancing. It is not possible to track one individual across generations, but it is possible to track ancestries. Each ancestry is assigned a number. As you move down the matrix individuals of the same ancestries can be tracked through time. In each row the ancestries are always displayed in ascending order from left to right.

The second matrix is the colour map matrix. The colour map matrix

is an m x 3 matrix where m is number of colours in the image. The three columns contain numbers from 0 to 1 representing the amount of red, green, and blue, respectively. There is one row for each unique colour to represent an ancestry in the image matrix. It is not simple to generate an infinite number of colours that are visually distinguishable. Therefore, if there are more than 7 ancestries to be tracked the colours will repeat in our modular colour mapping scheme. In the image the ancestries of the same colour can be distinguished as long as some individuals from the 6 intervening ancestries remain.

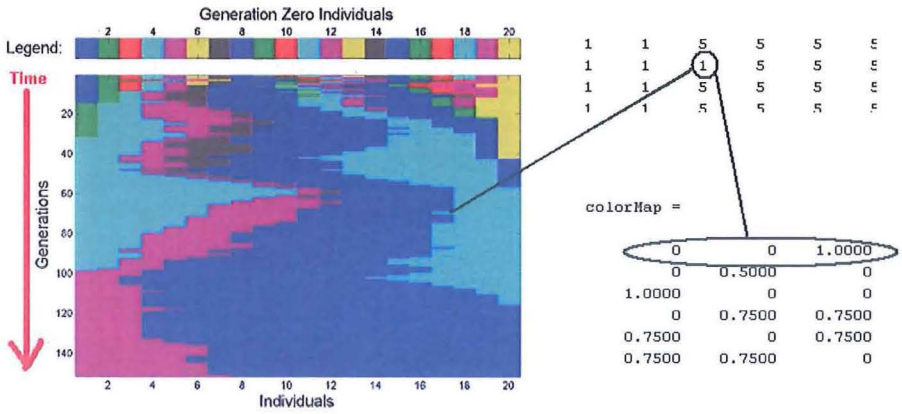


Figure 2: Sample Visualization

5 Population Models

The Coalescent process holds for a variety of population models. All models must fit the following three assumptions:

1. Genetic differences have no consequence of fitness.

2. The population is not subdivided, geographically or otherwise.

3. The size of the population is constant over time.[9, pg. 43]

We study two such models that are popular in population genetics. The fictitious organisms in our models are extremely simplified when compared to creatures in nature. In fact, there are no creatures in nature that have the DNA structure and reproductive methods we will use. Unless stated otherwise, the organisms are haploid, reproduction is asexual and the population size, N, is fixed.

5.1 Moran Model

The Moran Model is characterized by overlapping generations[2, pg 262]. In each generation there is one birth and one death. One member of the prior generation is randomly selected for reproduction, another is randomly selected for death. The current generation is then constructed with all the individuals from the prior generation, except for the individual selected for death and a copy of the individual selected for reproduction. If the individual selected for death is the same as the individual selected for reproduction, then the current generation is an exact copy of the prior generation. If different individuals and selected for birth and death then, the reproduced individual replaces the dead individual.

To visualize the Moran we will track each individual separately (each individual is assigned a different allele). A 1 x N population matrix is built to hold the starting population. This population is then bred a fixed number of times according to the Moran Model. During each breeding we take two

permutations of the prior generation. The first element in the first permutation is selected for birth. The first element in the second permutation is select for death. The population matrix is updated by replacing the value in the death positions with the value in the birth position. A sorted copy of the population matrix is then appended to the visualization matrix.

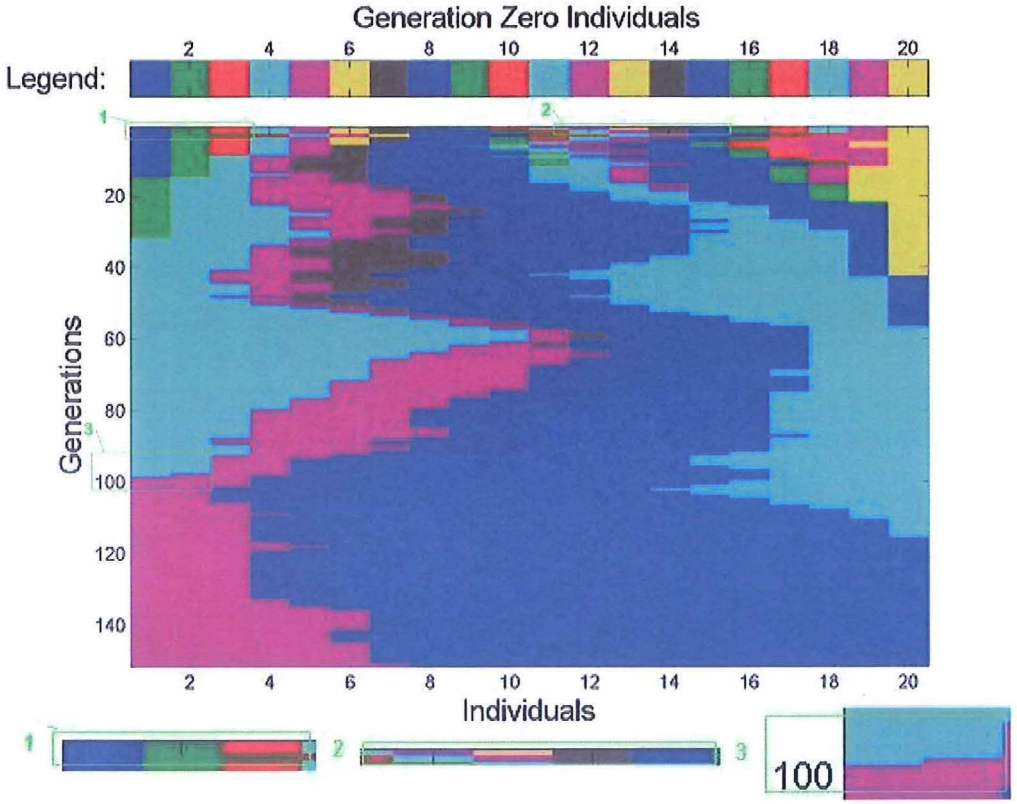


Figure 3: Moran visualization for 20 individuals over 150 generations

Looking at the example in Figure 3 we see that only a handful of ancestries dominate after several hundred generations. Running the visualization several times will allow us to see different individuals dominate each time. It is worthwhile to reiterate that only genetic drift selection is occurring. Con-

centrating on the first breeding event of this population - individual 2 was selected for birth and individual 13 was selected for death. In generations ninety eight through one hundred (inset 3) we see individual four go from two ancestors, to one, to none. Since all succeeding generations are drawn from the current, even if the model was run forever, we would never see another individual from ancestry 4 appear.

5.2 Wright-Fisher Model

The Write-Fisher model belongs to a larger class of models called the Cannings Models it is characterized by complete replacement of all individuals in each generation[2]. In order for the total population size(N) to remain the same, N reproduction events are selected. Each reproduction event is a random selection of the individuals of the prior generation. This means that an individual can have multiple offspring and that a given individual in the current generation has an equal probability of having each member of the previous generation as its parent.

To visualize the Wright-Fisher model we follow the same basic outline as the Moran Model. We again track each individual separately (each individual is assigned a different allele). An 1 x N population matrix is built to hold the starting population. The population is then bred according to the Write-Fisher model. During each breeding cycle we select a number uniformly at random from the set $\{1, 2, \ldots, N\}$ N times. Each one of these N numbers represents a position in the population matrix. We construct the next generation with the same ancestry as the individuals that occupy those

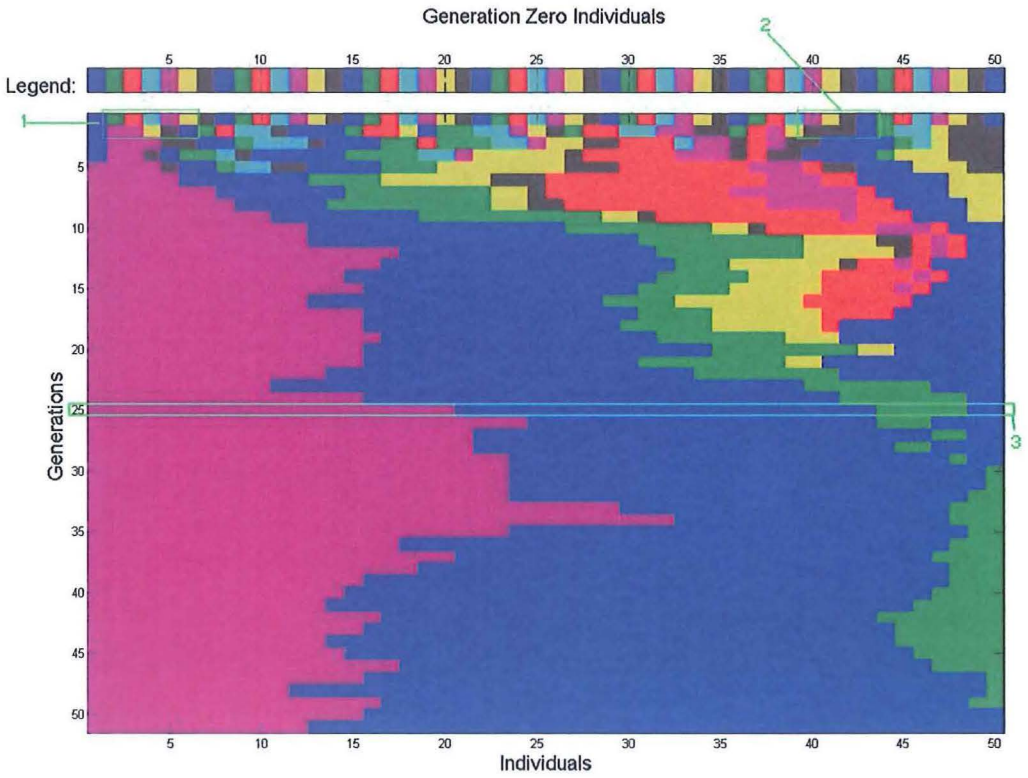


Figure 4: Wright-Fisher visualization for 50 individuals over 50 generations

N positions in the prior generation. As before a sorted copy of the population matrix is then appended to the visualization matrix.

Looking at the example in Figure 4 and comparing it to Figure 3 we see that change is much swifter in the the Write-Fisher model. Ancestries two, three, and four have all died out at generation one (inset 1). While ancestry 42 now has three individuals(inset 2). After twenty-five generations only three ancestries of the original 25 ancestries remain(inset 3).

6 Genetic Drift

"Genetic drift is the process in which diversity is lost from a population[1]". This loss in diversity is due to "the random process of genetic transmission and of birth and death of individuals within populations[9, pg. 4]." In the absence of mutations and recombinations to introduce diversity all individuals will eventually share the allele of a single common ancestor[1].

To make the genetic drift measure in the Moran Model easily comparable to genetic drift measure in the Wright-Fisher model we introduce a new time measure into the Moran Model called a step. Because there is at most a small change in each breeding event and all N individuals have an equal chance to change in each breeding event, we will redefine 1 breeding event in the Moran model as a step and redefine a generation as N steps[9, pg. 43].

To create the genetic drift visualization using the Moran Model we will no longer identify each individual with a different allele. Instead we will split the individuals between two allele types. The first allele is represented by 1, the second by 2. Each individual in the population is randomly assigned to one of the two allelic groups. The population is then bred according to the Moran Model. At the end of each row in the image matrix a summary column is added. The summary contains the same percentage of each colour as the population. The summary is re-calculated after each generation (N steps). The legend on the right shows 100 possible summary colours. The simulation stops when all individuals have the same allele.

Looking at Figure 5 we see that in generation zero twelve individuals are of allele type 1 and thirteen are of allele type 2. In the neighborhood of

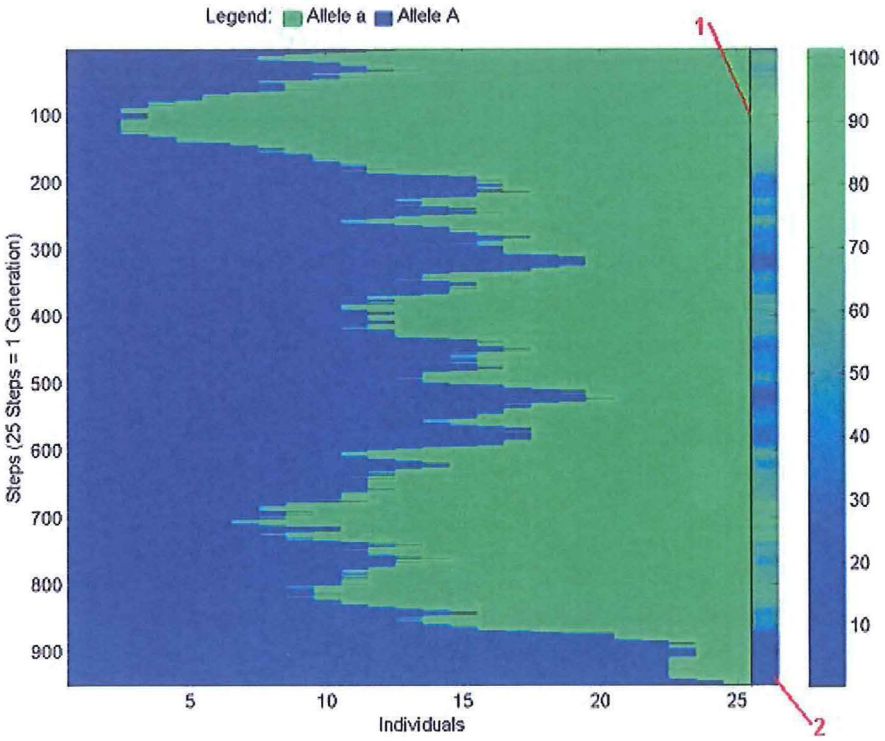


Figure 5: Moran visualization for 25 individuals with 2 Alleles

generation 100, we can see that when all individuals except one have the same allele the outcome of the trial is not assured(inset 1). Only when all individuals have the same allele can we stop the simulation(inset 2).

To build the Wright-Fisher visualization of genetic drift we use the same process as the Moran visualization of genetic drift, assigning each member of the population to one of two allele types and summarizing each generation in the final right-hand column. The only difference from the Moran model is the use of the Wright-Fisher breeding scheme.

Reviewing the sample output of the Wright-Fisher visualization in Figure 6 we see that the shifts take place more quickly. This is the reason that the generation is redefined in the Moran Model to incorporate N breeding

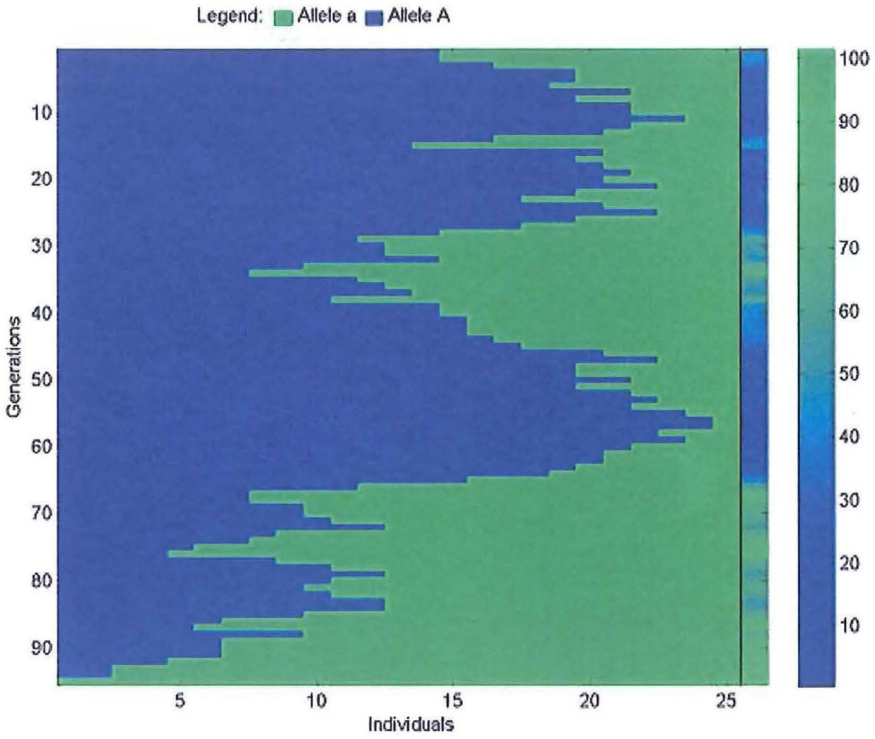


Figure 6: Wright-Fisher visualization for 25 individuals with 2 Alleles

events. This does not mean that all Wright-Fisher simulations will be shorter than all Moran simulations. We are still dealing with a stochastic process, therefore, a given simulation of the Wright-Fisher could be longer that a given simulation of the Moran. On average we expect the Wright-Fisher simulation to be N times shorter that the Moran. Taking the rescaling into account, the structure is the same between Moran and Wright-Fisher. The differing models are not affecting the structure of the stochastic process.

7 Reversing Time

Up to this point we have been running our models forward in time. We will now be looking back in time. We begin by once again looking at a forward time model of Moran and discussing the hurdles we will face if we tried to reconstruct it in reverse.

We begin with the left most picture in Figure 7. It is a forward simulation using the Moran model with full population information, ie. tracking the ancestry of all individuals from generation zero. This picture contains full information about the ancestry of all individuals. What information would we not have access to if we were working backwards? Firstly, the current generation only contains individuals from ancestries 5 and 16. We would not know about the shape or structure of any other ancestries. The middle picture shows the forward simulation with all of the other ancestries removed. Secondly, we would not know about any individuals in ancestries 5 and 16 that do not have direct descendants in the current generation. As we move back in time the number of individuals in a generation has to be less than or equal to the number in generations succeeding it. In the right picture of Figure 7 the individuals that outnumber the succeeding generations are marked in light tan and light red. These individuals can not have direct descendants in the last generation. Finally, we would name the ancestries 1 and 2 because they are the first and second ancestries in the current generation. The individuals marked in dark red and dark tan in the right most picture represent the best case we could hope for if we did a reverse time model of the picture on the left. Figure 7 helps us see how much information we are

missing when we do reverse time simulations.

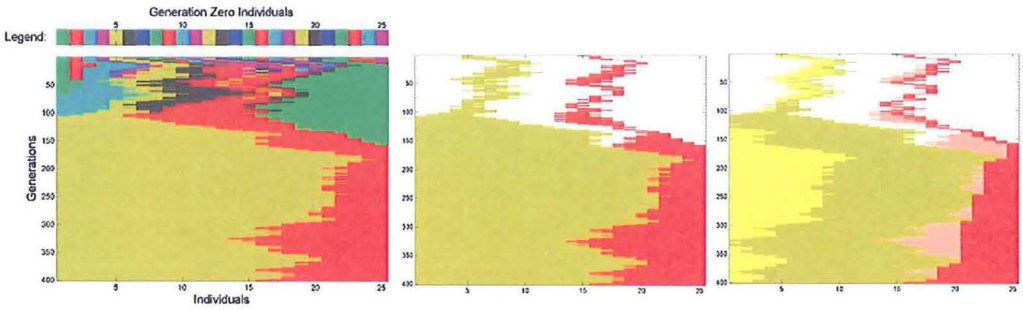


Figure 7: Comparison of Information Available

To construct a visualization of the Moran Model going back in time we start with a 1 x N matrix to hold the population. Then we apply the reverse breeding step. The reverse breeding step contains an 'unbirth' and an 'undeath'. The 'unbirth' is accomplished by randomly selecting a number, b, from 1,2....N. The 'undeath' is accomplished by randomly selecting a number, d, from 1,2,...N. If the two numbers b and d differ, the individual in position b is replaced with a place holder (white space). If b equals d, then nothing changes. The possibility of b equaling d is $1/N^2$, this is the probability in the forward model that the some individual in selected for birth and death. There is an additional check made to see if the individual selected for deletion is the last individual of an allele type. The last individual of an allele type can not be removed because in order for it to survive in the current population one of its ancestors must have existed at all times in the past. If the last individual of an allele type is selected, the breeding trial is discarded and another breeding trial is begun.

The two samples in Figure 8 represent only having genetic data for ten

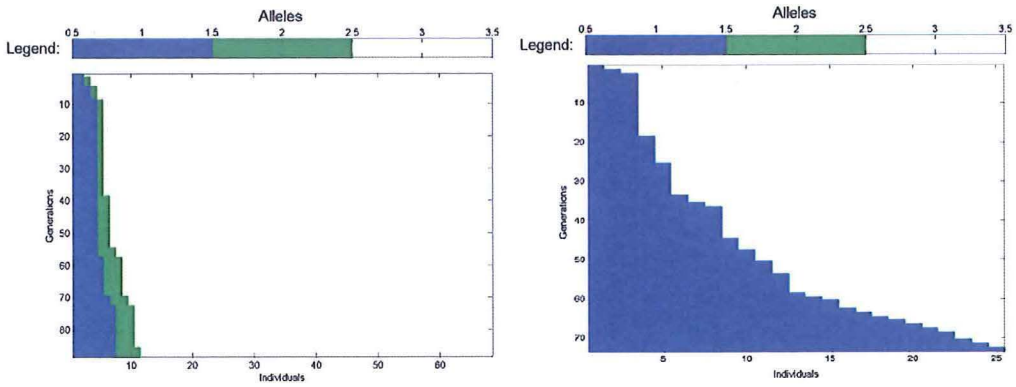


Figure 8: Reverse Wright-Fisher Visualizations

people in a population of seventy, having a population of twenty-five individuals of one allele type and having full genetic data. In looking at the samples of the Moran Model in reverse we can see that the more white-space there is in a generation the longer it takes for the population to change. Because we have not introduced mutation eventually all our realizations are reduced to one individual from each allele group.

To construct a visualization of the Write-Fisher Model going back in time we start with a 1 x N matrix to hold the population. Then we apply the reverse breeding step. In the Wright-Fisher model each individual selects a parent uniformly at random from the previous generation. To simulate this we select a number from the set $\{1, 2, \ldots, N\}$ N times. We call this matrix the paternity matrix, each of the N numbers in it represents the position of the parent in the previous generation. We now construct the last generation by incrementing through the N positions and looking at the paternity matrix to see who has chosen this position. If multiple individuals of the same type select the same parent, then the parent is assigned to that type. If some

children are of unknown type and the rest are of the same type, then the parent is assigned the known type. If no one selects a parent, then it is assigned an unknown type. If two individuals of different allelic types select the same parent, then the breeding cycle is aborted.

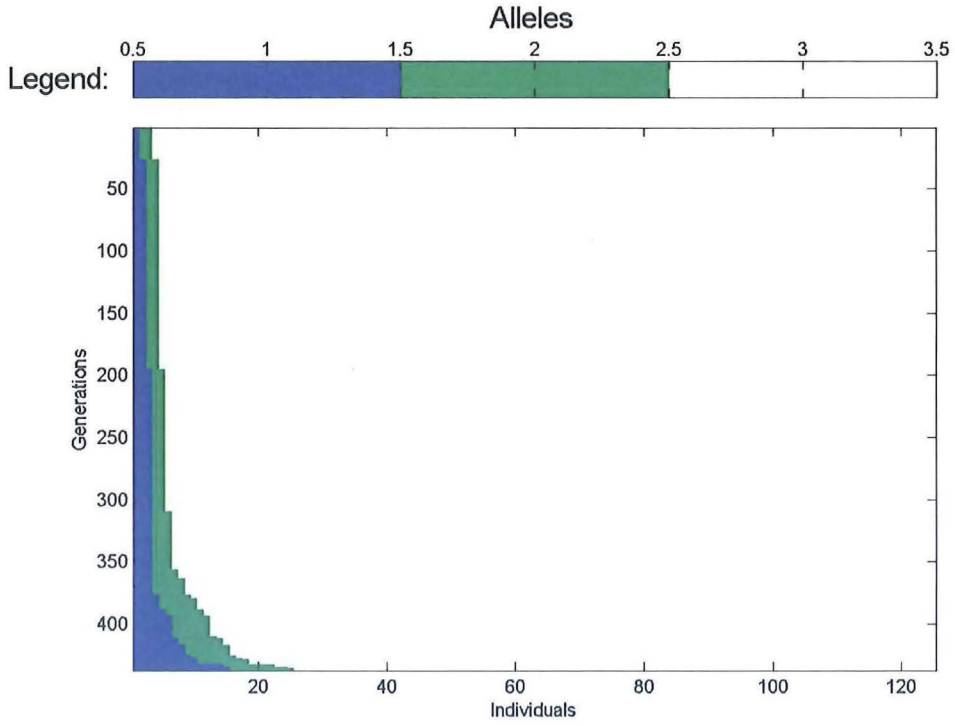


Figure 9: Reverse Wright-Fisher Visualization for 22 individuals from a population of 122

The sample in Figure 9 represents a population of 122 where we know the allelic type of twelve individuals. Just as we saw in genetic drift, the breeding process does not affect the stochastic structure. The Write-Fisher pictures are the same shape as the Moran pictures.

8 Coalescent Time Scale

"In order to obtain Kingman's coalescent process in the Wright-Fisher model and the Moran model" we must rescale time[9, pg. 50]. Coalescent time only counts the points where the population as a whole changes. "Each coalescent event decreases the ancestral lineage by one[9, pg. 43]." Because our ultimate goal is to find the MRCA of the entire population, the population "defines a natural time unit of N generations[4]."

When we have full population information as we did in the early forward time simulations the population changes at every time step. In the reverse models, especially those where we only know a small fraction of the individual's genetic information we can see large time periods between population level changes. The further back in time we go, the less information we have about the population and the longer it takes for a coalescent event to occur. "The most ancient coalescent time, the one in which the remaining two lineages coalesce into the MRCA should be the longest[9, pg. 45]." Kingman showed that as population size approaches infinity coalescent times are independent and exponentially distributed[4, pg. 37].

To build the Moran Coalescent and the Write-Fisher Coalescent simulation we simple adjust the time scale on the backward simulation and add a black line to the image matrix each time the population changes.

Figure 10 is a backward simulation for a population of eight where we have full genetic information for the population. This is a very small population, but it allows us to see the steps and how the black lines are inserted between each change. Figure 11 is of a population of 110 where we have genetic

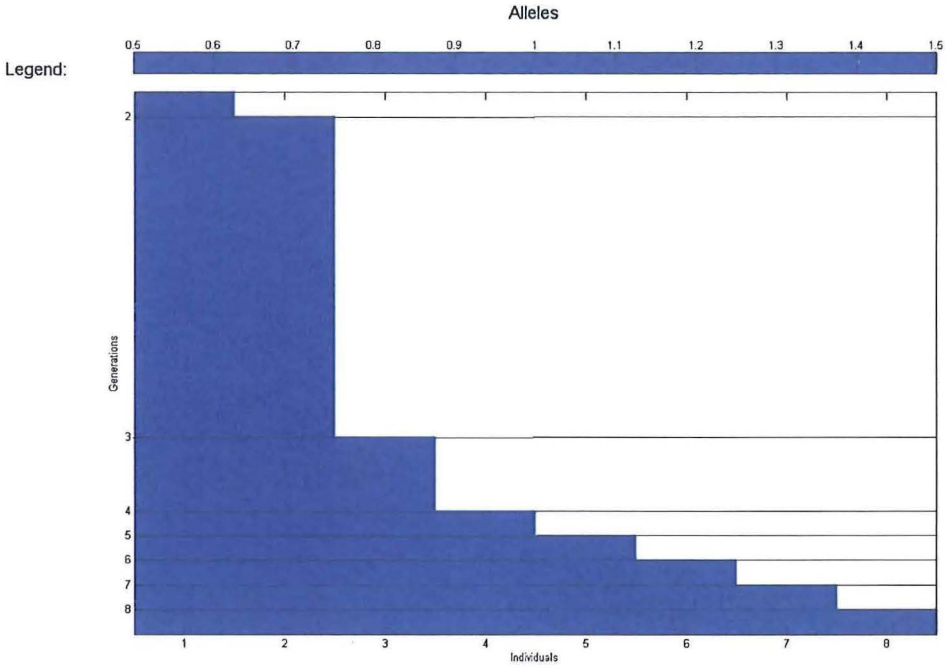


Figure 10: Moran visualization for 8 individuals over 8 Coalescent Time Steps

information from 10 individuals. This is still a very small population, but we can start to see that coalescent time steps are larger further back in time.

9 Conclusion

"The circle of ideas that has come to be known as the coalescent has proved to be a useful tool in a range of genetical problems, both in modeling biological phenomena and in making statistical sense of the rich data now available[5]." The amount of information available is growing rapidly. The human genome has been mapped. In April 2003 a working draft of the sequence of the human genome was published in Science by the Human Genome Project[7]. "The International HapMap Project is a multi-country effort to identify and

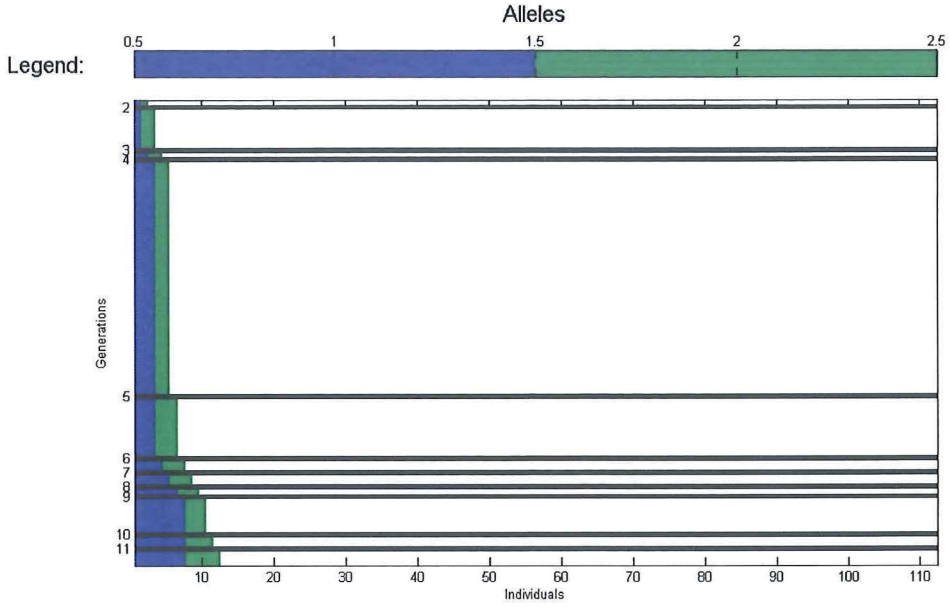


Figure 11: Wright-Fisher visualization with coalescent time scale for 110 individuals genetic information for 10 individuals

catalog genetic similarities and differences in human beings[8]."

In order for population geneticists to apply the tools that build on the coalescent, they must first understand the underlying models and assumptions. Statistical geneticists are tasked with communicating the results of their work in a way that non-statisticians can understand. Visualization can play a key role in this communication.

Kingman, the father of Coalescent Theory, put it very plainly when he said, "Those who analyze stochastic models should always lift their eyes from their equations to ask what they actually mean[5]."

References

[1] Ovide Arino Adam Bobrowski, Marek Kimmel and Ranajit Chakraborty. A semigroup representation and asymptotic behavior of certian statistics of the fisher-wright-moran coalescent. *Handbook of Statistics*, 19:215–248, 2001.

[2] C. Cannings. The latent roots of certain markov chains arising in genetics: A new approach, i. haploid models. *Advances in Applied Probability*, 6(2):260–290, jun 1974.

[3] Brian S. Everitt and Graham Dunn. *Applied Multivariate Analysis.* Arnold, 2001.

[4] J. F. C. Kingman. On the genealogy of large populations. *Journal of Applied Probability*, 19:27–43, 1982.

[5] J. F. C. Kingman. Origins of the Coalescent: 1974-1982. *Genetics*, 156(4):1461–1463, 2000.

[6] John W. Tukey. Data-based graphics: Visual display in the decades to come. *Statistical Science*, 5(3):327–339, aug 1990.

[7] Unknown. The human genome project. Website, 6-Feb-2008. `http://www.ornl.gov/sci/techresources/Human_Genome/project/about.shtml`.

[8] Unknown. International hapmap project. Website, 6-Feb-2008. `http://www.hapmap.org/thehapmap.html.en`.

[9] John Wakely. *Coalescent Theory, An Introduction.* Unpublished.

Appendix A - Code Excepts

The majority of the code base is related to formatting the images. This code does not convey anything about the models themselves. In contrast the breeding portions of the code convey the mechanisms of the models concisely. Before reviewing the breeding functions we will explains the input and output arguments.

A.1 Input and Output Variables

```
N           % initial population size
pop         % population matrix of length N
fill        % correspond to white in the colour map
trouble     % if trouble = 1, ignore the results
```

A.2 Moran Breeding Step

```
function pop = moran_breed(N,pop,fill)
    b = randperm(N); %select 1 birth (do perm of pop, select 1st)
    b=b(1);
    d = randperm(N); %select 1 death (do perm of pop, select 1st)
    d = d(1);
    pop(d) = pop(b); %replace dead with just born
end
```

A.3 Wright-Fisher Breeding Step

```
function pop = wf_breed(N,pop,fill)
    pop=pop(ceil(rand(1,N)*N)); % Wright-Fisher Model Breeding
end
```

A.4 Moran Reverse Breeding Step

```
function [pop trouble] = moran_breed_rev(N,pop,fill)
    trouble = 1
    b = randperm(N); %select 1 birth (do perm of pop, select 1st)
    b = b(1);
    d = randperm(N); %select second number.
                        %It will match the first 1/N^2 times.
                        %When the happens do nothing.
    d = d(1);
    if sum(pop==pop(b)) > 1 % drawn again if you select a solo ind.
                                %no lineage can die out.
        if d ~= b % chance that no one eliminated.
            pop(b) = fill;
        end
        trouble = 0;
    else
        trouble = 1;
    end
end
```

A.5 Wright-Fisher Reverse Breeding Step

```
function [pop trouble] = wf_breed_rev(N,pop,fill)
    parents=ceil(rand(1,N)*N); % Wright-Fisher Model
    % each ind picks their parent
    % make sure that 2 ind with different allele
    % type didn't pick the same parent,
    % check each and update population.
    % Abandon if there is any problem
    new_pop = ones(1,N) * fill; %seed new gen with fill values
    trouble = 0;
    for i=1:length(pop)
        % get all the offspring for this ind
        child = unique(pop(parents==i));
        %if there are no children remove [0] (leaves [] of length 0)
        if sum(child==0) > 0, child = setdiff(child, [0]); end
        %if there are filler children remove and disregard
        if sum(child==fill) > 0, child = setdiff(child, [fill]); end
        if length(child) > 1
            trouble = 1;
            break;
        else
            if length(child) > 0, new_pop(i) = child; end
        end
    end if trouble == 0
```

```
        pop = [new_pop];
    end
end
```